A Perception and PDE Based Nonlinear Transformation for Processing Spoken Words

Yingyong Qi and Jack Xin

Abstract

A perception and PDE (partial differential equation) based nonlinear transformation is proposed to process spoken words in noisy environment. The transformation was designed to reduce noise through time adaptation and spectral enhancement by evolving a focusing quadratic fourth order nonlinear PDE (the Cahn-Hillard equation). Constant, low SNR signals were adaptively decreased to reduce noise. Once the noise levels were reduced, the evolution of focusing PDE was used to enhance the spectral peaks, and further reduce noise interference. Numerical results on noisy spoken words indicated that the transformed spectral pattern of the spoken words was insensitive to noise. The spectral distances between noisy words and original words were significantly reduced after the transformation.

1 Introduction

For speech signals in a noisy environment, the spectrogram contains multiple scale information, and the essential features of speech signal (e.g., formant frequencies) are buried inside a masking noise. Thus, extracting speech spectral features and suppress noise is one necessary step in speech recognition. It is well-known that incorporating time adaptation, isolating spectral peaks and enhancing peak magnitudes, help the recognition task to focus on phonetically meaningful aspects of spectrogram, and reduce the sensitivity of the pattern recognition to noise interference. Convincing experimental evidence to these effects have been demonstrated by Strobe and Alwan [6], Hermansky and Morgan [4], among others.

However, the reported methods on signal processing involve many ad hoc procedures, and proceed in a more or less case by case basis. Our objective here is to introduce a systematic mathematical approach to perform adaptation and peak isolation (focusing), based on both human hearing perception and properties of a class of nonlinear PDE’s (the focusing Cahn-Hillard equation with quadratic nonlinearities). The end result of our treatment is a nonlinear transformation that distinguishes main speech spectral features and removes the noisy disturbances.

2 Nonlinear Transformation, Algorithms and Properties

Let \( a_{i,j} = a(t_i, x_j), 1 \leq i \leq I, 1 \leq j \leq J \), be the raw speech spectral data arising from short-time window Fourier transform of a sound wave. Let the log-magnitude power spectrum be \( u_{i,j} = \log|a_{i,j}|^2 \). We shall place the \( I \times J \) points uniformly on a rectangular domain \([0, T] \times [0, X]\) so that the grid size is \( h \in (0, 0.09) \). When noise is present, \( u_{i,j} \) receives a lot of energy towards the high frequency region. Noise also makes \( u_{i,j} \)
Time and frequency directions must be treated differently, because speech signals are strongly time dependent and the time window for processing is short. In human auditory processing, if the spectral amplitude at any fixed frequency is not varied enough over a time window, the human auditory system is going to reduce its response and ignore that segment of signal. This phenomenon is called adaptation in time and is related to the decay of an onset response of a neuron with time.

We used $\delta_{k,j}$ as a measure of signal caused variation. If $\delta_{k,j}$ is below a threshold value in $dB_c$ ($dB_c = 2$ in our calculation), the adaptation in time will ensue. The adaptation processing is done continuously by a discrete (e.g. finite difference) approximation of the nonlinear local PDE ($\delta(u)$ viewed as a nonlocal operator in $x$):

$$u_t + f(\delta(u) - dB_c, \tau)u = 0,$$

where $f$ is a nonnegative nonlinear function, monotone decreasing in $\tau$, the processing time, and behaves as a transition layer (a step function) in $\delta(u) - dB_c$. One simple choice of $f$ is:

$$f = \gamma \chi_{\{\delta(u) - dB_c \leq 0\}} \cdot \chi_{\{|\tau| < \tau_0\}},$$

where $\chi$ is the characteristic function, $\gamma > 0$, the adaptation rate, and $\tau_0 > 0$ the time scale of adaptation. Notice that if $\delta(u)$ is above the threshold $dB_c$ or if $u = 0$, the solution will reach a steady state.

For small $\tau$, solution $u$ to (2.1) is well approximated by:

$$u_0 \exp\{-\gamma \chi_{\{\delta(u_0) - dB_c \leq 0\}} \tau\}.$$  

Subsequent to adaptation comes the peak isolation and peak enhancement stage of processing. A PDE that is able to locally focus peaks (large curvature regions and not flat regions), and preserve $L^\infty$ norm (or energy on the log scale magnitude), will serve the purpose. One such PDE is the following quadratic focusing Cahn-Hillard (C-H) equation:

$$u_t = -\alpha (u^2)_{xx} - eu_{xxxx},$$

where $\alpha \geq 1$, $0 < \epsilon \ll 1$. The equation (2.4) can be rewritten as:

$$u_t + 2\alpha u_x u_x = -2\alpha u_{xx} - eu_{xxxx},$$

hence positive curvature regions ($u_{xx} < 0$) go up, negative curvature regions ($u_{xx} > 0$) go down. Moreover, positive slope regions advects to the right, and negative slope regions advects to the left. Combining these two effects, we have the desired local peak focusing. This motion is however highly unstable, and requires a little bit of high wave number stabilization ($\epsilon > 0$) to be numerically computable, thus the C-H equation (2.4).

To summarize, our nonlinear transformation consists of (1) time adaptation using perception based three frequency channel envelopes and (2) local spectral peak focusing using the Cahn-Hillard equation (2.4). The transformation will be shown via numerical examples to yield stable results under noise perturbations, and provides useful information for recognition of noisy data.

3 Numerical Results and Relations to Other Processing Methods

The method is tested on spoken digits including zero and oh for SNR = 0, 3, 5, 10, 20 dB. Results indicated that The $L^2$ spectral distances between processed signals and the corresponding noise free signals are reduced by as much as 50% for SNR = 3, 5, 10 dB. Numerical experiment at SNR = 5 dB shows 100% recognition rate on the 11 spoken digits.

Let us comment on how our transformation is related to previous works. The RASTA (relative spectra) method of [4] exploits the different temporal spectral behavior of noise and signals. Time adaptation is a key component of the method in that RASTA suppressed constant or less varied portions in the time trajectory of each frequency before applying the usual rational approximation (all pole modeling). Besides this,
RASTA also transforms the amplitude nonlinearly based on human hearing laws. The Fig. 8 and advocated in [6]. In the above referenced works, the processing used estimation methods such as all pole modeling, and raised-sine cepstral liftering technique for peak isolation. The basic idea of cepstral liftering is to expand the even log spectrum in terms of cosine basis functions, with the first few coefficients \((c_0, c_1, \text{ and higher})\) representing log-spectrum average \((c_0)\), log-spectrum tilt \((c_1)\) and varying ripples \((c_1)\). Then the so called raised-sine lifter deemphasizes slow changes with frequency \((\text{related to overall level and vocal driving-function characteristics})\) as well as fast changes \((\text{numerically})\). The processing emphasizes moderate variations with frequency, and so both spectral peaks and valleys.

Our method is based on properties of solutions to the focusing Cahn-Hillard equations. It treats the raw data directly without all pole approximation or discrete cosine transform, and offers more flexibility and robustness in handling peak enhancement because it uses local geometrical information \((\text{curvature, slopes})\) without computing the expansion coefficients which also may well encode undesired features from other points \((\text{such as ripples and roughness from other frequencies away from the harmonics})\).

### 4 Analytical Properties of the Focusing Cahn-Hillard

The C-H equation (2.4) and related fourth order parabolic equations are known to be well-posed locally in time in suitable Sobolev spaces, [3],[2] among others. When (2.4) is considered on the interval \([0, L]\) under the zero Neumann boundary conditions: \(u_t|_{x=0, L} = 0, \ u_{xx}|_{x=0, L} = 0, \) and initial condition: \(u_0(x) \in H^2([0, L])\) with zero trace \(u_{0,xx}|_{x=0, L} = 0, \) Elliot and Zheng [3] showed that if \(\epsilon > L^2/\pi^2, \) and \(u_0\) is sufficiently small, then there is a global solution \(u \in H^1([0, \infty); H^2([0, L])) \) so that:

\[
\lim_{\tau \to \infty} \|u - L^{-1} \int_0^\tau u(x) \, dx \|_{\infty} = \lim_{\tau \to \infty} \left( \|u_x\|_\infty + \|u_{xx}\|_2 \right) = 0. \tag{4.1}
\]

This result demonstrates the smoothing effect of the fourth order term.

If \(\epsilon\) is small or initial data is not small, solutions can blow up in finite time due to the quadratic term. We are more interested in the stage prior to blow up, that is the focusing regime, best illustrated on the entire line by self-similar solutions, which may serve as local approximations of focusing time dependent solutions.

With no loss of generality, set \(\alpha = 1, \) and \(u = (\tau^* - \tau)^{-1/2} f(x/(\tau^* - \tau)^{1/4}) \). Then it follows from (2.4) that \((\xi = x/(\tau^* - \tau)^{1/4}, \ t = \frac{d}{d\xi})\):

\[
\frac{1}{2} f + \frac{1}{4} \xi f' + \epsilon f'' + (f^2)'' = 0. \tag{4.2}
\]

We shall only consider even solution of (4.2), and so it is enough to let \(\xi \geq 0. \) Equation (4.2) has an unbounded solution \(f = -\frac{1}{12} \xi^2, \) on which the fourth derivative term vanishes identically. The corresponding solution \(u = -\frac{1}{12} \xi^2 \) shows the parabolic shape of blowup profile.

Equation (4.2) also has a first integral (Bernoff and Bertozzi [2]):

\[
\frac{1}{4} \xi^2 f - \frac{f^2}{2} + \xi(f^2)\xi + \xi f_{\xi\xi\xi} - f_{\xi} = c_0. \tag{4.3}
\]

Integrating (4.2) over \(\xi \geq 0\) indicates that any spatially decaying solution must change sign and has zero integral. Let us look for solutions to (4.2) such that \(f'(0) = f''(0) = 0, f(0) > 0, \) and \(f''(0) < 0. \) Let \(G = \int_0^\xi f(\xi) \, d\xi, \) then integrating (4.2) from zero to \(\xi \) gives:

\[
\frac{1}{4} G + \frac{1}{4} \xi G_{\xi} = -(f^2)\xi - f_{\xi\xi\xi},
\]

and using (4.3) to get:

\[
\frac{1}{4} G + \frac{1}{4} \xi^2 G_{\xi} = \frac{1}{4} \xi^2 G_{\xi} - (G_{\xi})^2 - G_{\xi\xi\xi} - c_0,
\]

3
In terms of $G$, any spatially decaying solution in $f$ corresponds to a positive solution $(G(\xi) > 0$ when $\xi > 0$) of (4.4) such that $G \sim 4c_0\xi^{-1} + h.o.t.$ for large $\xi$.

If we regard (4.4) as an initial value problem, with initial data: $G(0) = G^0(0) = 0$, $G'(0) > 0$, $G''(0) < 0$, $c_* = (G(0))^2 - G''(0) > 0$, then it is not hard to show that the solutions to the initial value problem increase to a local maximum from zero then go down. How they go down to near zero depend on $c_*$. The solutions are oscillatory about zero if $c_*$ is small when both $G(0)$ and $G''(0)$ are small in absolute value; while if $c_*$ is large, the solutions either blow up to negative infinity in finite time or become negative with a steep negative derivative. Hence the desired positive global solution with $G(\xi) \to 0$ can only be available for some bounded finite value of $c_*$. Numerical simulation of [2] showed such a solution and suggests that it exists uniquely and is dynamically attracting.

5 Conclusions

A perception and PDE based method is developed to process spoken words. Low, middle and high frequency channels are defined according to the critical bands of human audition. Difference of slowly varying energy envelopes are adaptatly decrease in time to reduce noise. A quadratic focusing Cahn-Hillard equation is evolved to enhance spectral peaks corresponding to fundamental voice frequencies. The method is tested on spoken words (spoken digits including zero and oh) for SNR = 0, 3, 5, 10, 20 dB. The $L^2$ spectral distances between processed signals and the corresponding noise free signals are reduced by as much as 50% for SNR= 3, 5, 10 dB. Numerical experiment at SNR = 5 dB shows 100% recognition rate on 11 spoken digits. More numerical experiments and human hearing based modeling will be carried out for other spoken words, and at even lower SNR values.

6 Acknowledgements

J. Xin would like to thank Profs. L. Caffarelli, T. Hou and S. Osher for helpful discussions.

References


